



16th THE GREEN 500

November 2014

Wu Feng

The Ultimate Goal of “The Green500 List”

- Raise awareness in the energy efficiency of supercomputing.
 - Drive energy efficiency as a first-order design constraint (on par with performance).

Encourage fair use of the list rankings to promote energy efficiency in high-performance computing systems.



Agenda

- Overview of the Green500 (Wu Feng)
- Methodologies for Measuring Power (Erich Strohmaier)
- Re-Visiting Power Measurement for the Green500 (Thomas Scogland)
- The 16th Green500 List (Wu Feng)
 - Trends and Evolution
 - Awards
- A Talk from #1 Supercomputer on the Green500

Brief History:

From Green Destiny to The *Green500* List

2/2002: Green Destiny (<http://sss.lanl.gov/> → <http://sss.cs.vt.edu/>)

- “Honey, I Shrunk the Beowulf!” 31st Int’l Conf. on Parallel Processing, August 2002.

4/2005: Workshop on High-Performance, Power-Aware Computing

- Keynote address generates initial discussion for *Green500* List

4/2006 and 9/2006: Making a Case for a *Green500* List

- Workshop on High-Performance, Power-Aware Computing
- Jack Dongarra’s CCGSC Workshop “The Final Push” (Dan Fay)

9/2006: Founding of *Green500*: Web Site and RFC (Chung-Hsing Hsu)

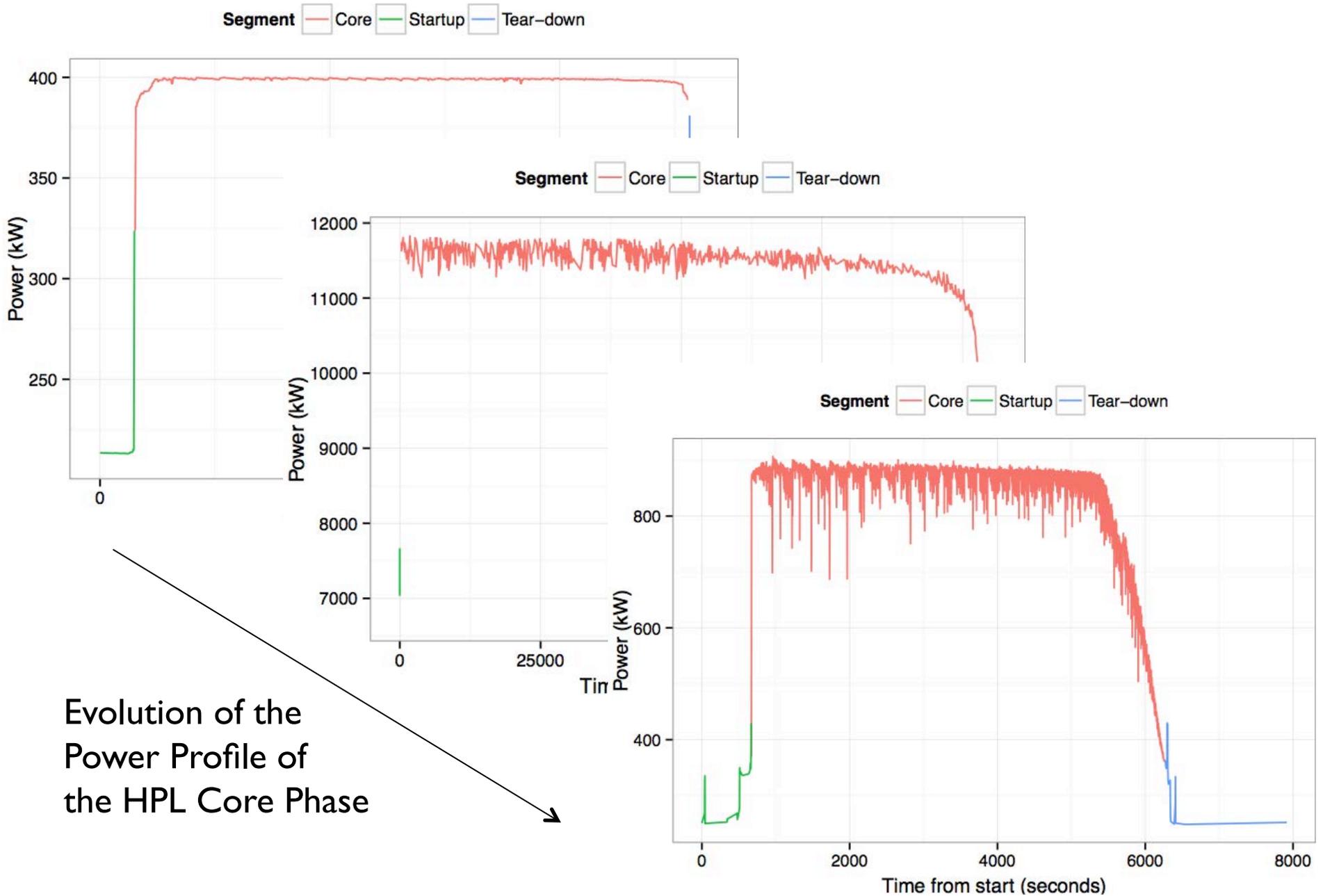
- <http://www.green500.org/> Generates feedback from hundreds

11/2007: Launch of the First *Green500* List (Kirk Cameron)

- <http://www.green500.org/lists/green200711>

Evolution of The Green500

- **11/2009:** Experimental Lists Created
 - *Little Green500:* More focus on LINPACK energy efficiency than on LINPACK performance in order to foster innovation
 - *HPCC Green500:* Alternative workload (i.e., HPC Challenge benchmarks) to evaluate energy efficiency
 - *Open Green500:* Enabling alternative innovative approaches for LINPACK to improve performance and energy efficiency, e.g., mixed precision
- **11/2010:** First Green500 Official Run Rules Released
- **11/2010:** Open Green500 Merged into Little Green500
- **06/2011:** Collaborations Begin on Methodologies for Measuring the Energy Efficiency of Supercomputers
- **06/2013:** Adoption of New Power Measurement Methodology (EE HPC WG, The Green Grid, Top500, and Green500)

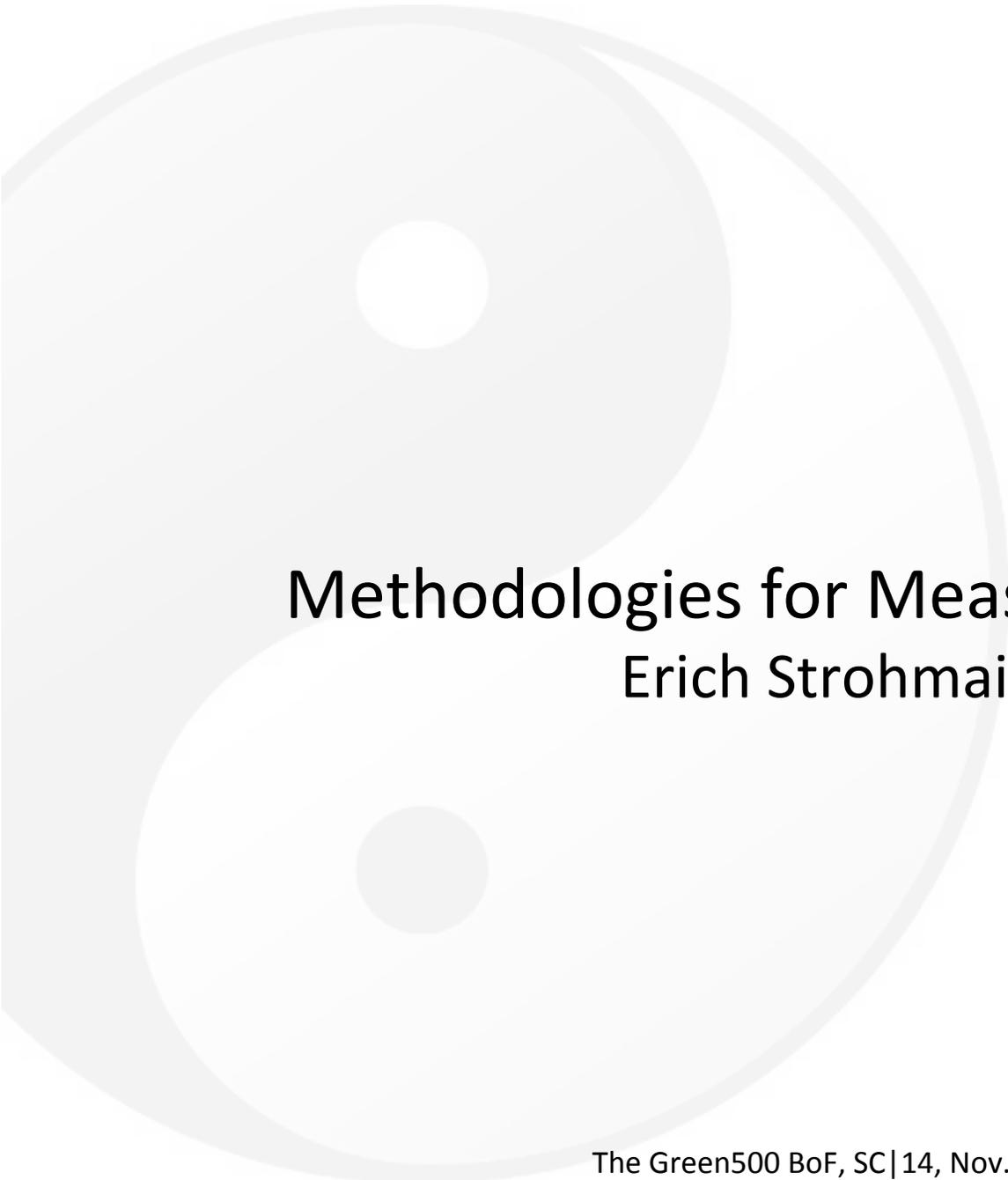


Evolution of the Power Profile of the HPL Core Phase

Legacy Assumptions

- Measuring a small part of a system and scaling it up does *not* introduce too much of an error
- The power draw of the interconnect fabric is *not* significant when compared to the compute system
- The workload phase of HPL will look similar on *all* HPC systems

These assumptions need to be re-visited.



Methodologies for Measuring Power

Erich Strohmaier

The Green500 BoF, SC|14, Nov. 2014
POC: info@green500.org

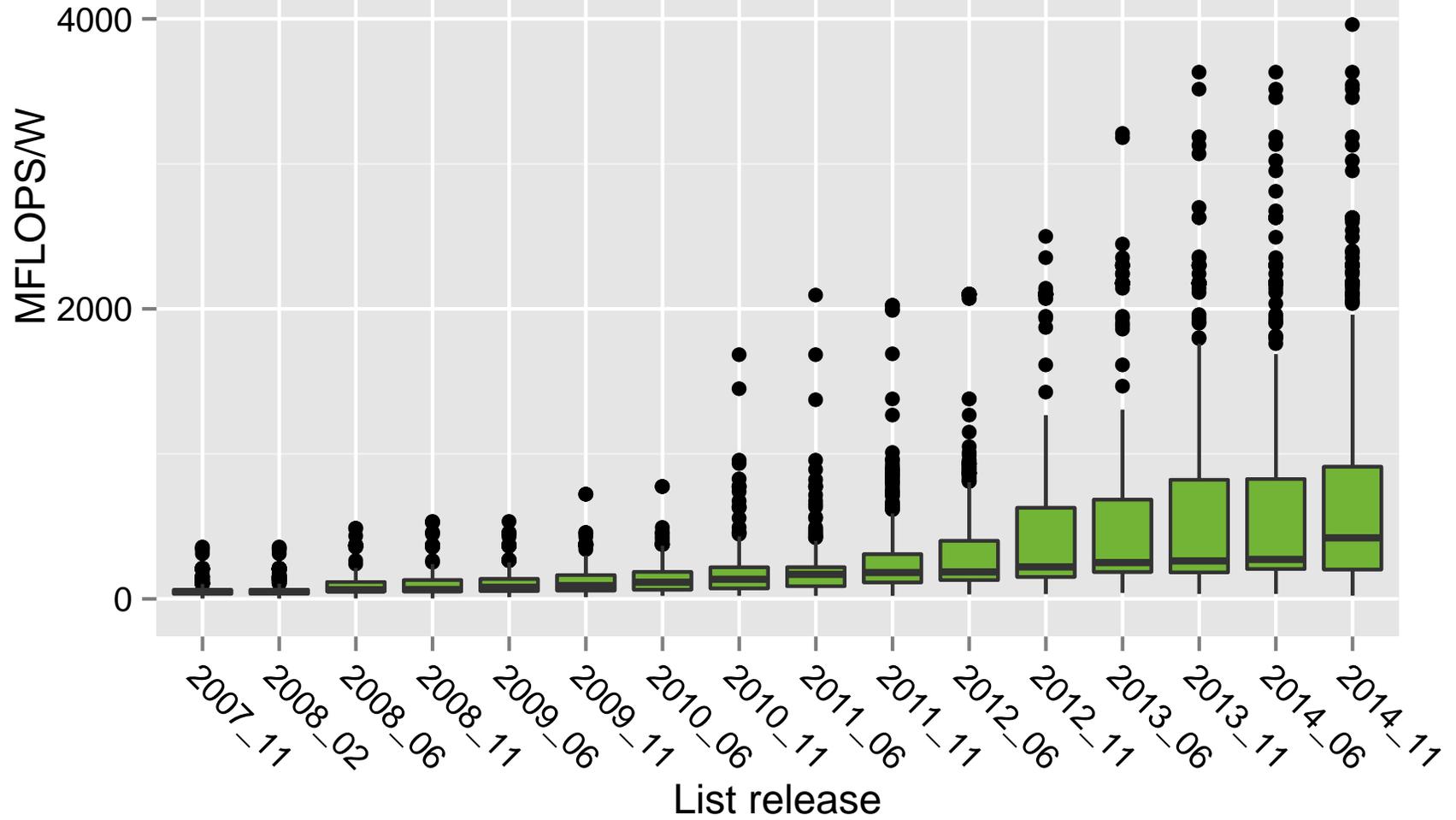


Re-Visiting Power Measurement for the Green500

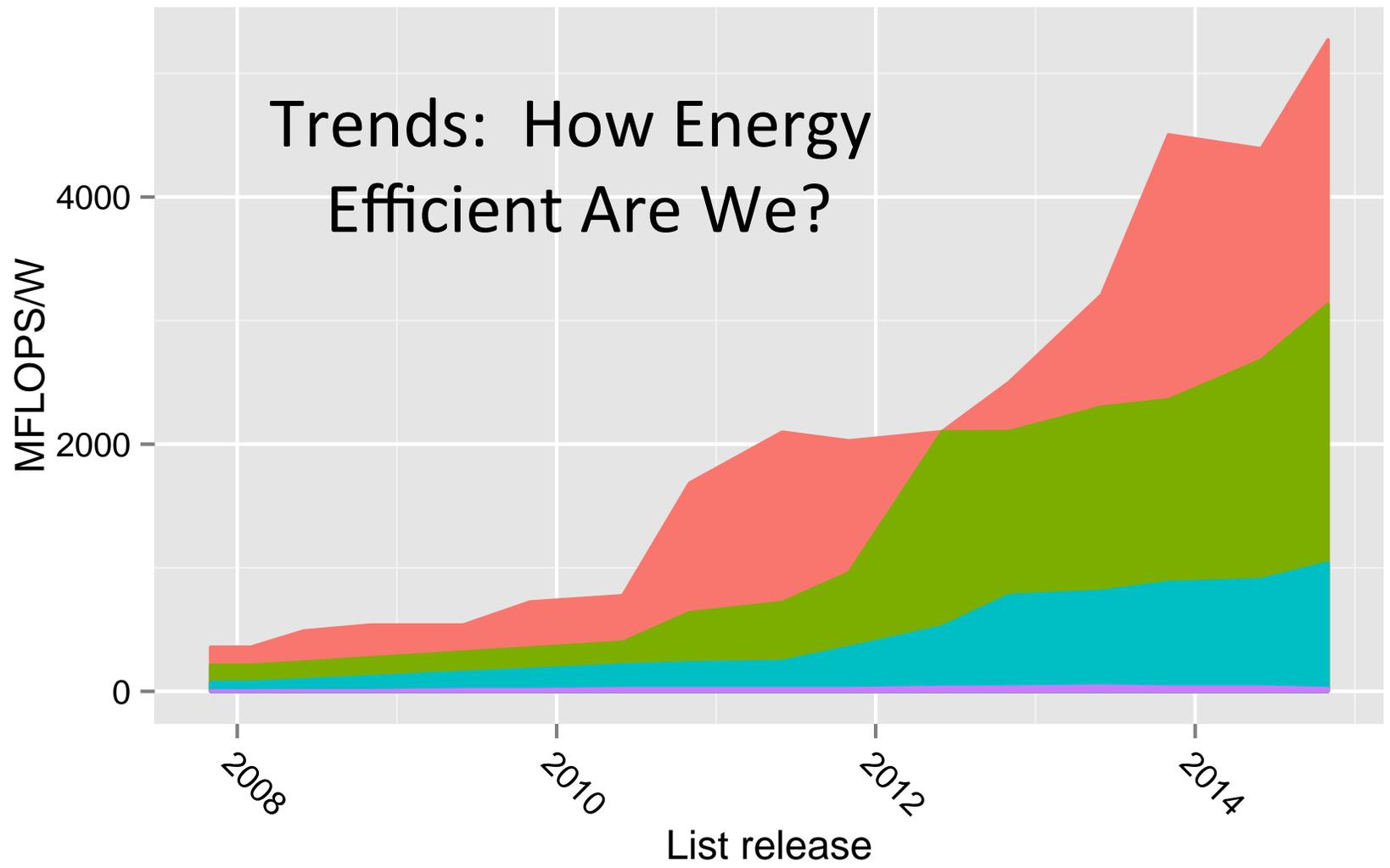
Tom Scogland

The Green500 BoF, SC|14, Nov. 2014
POC: info@green500.org

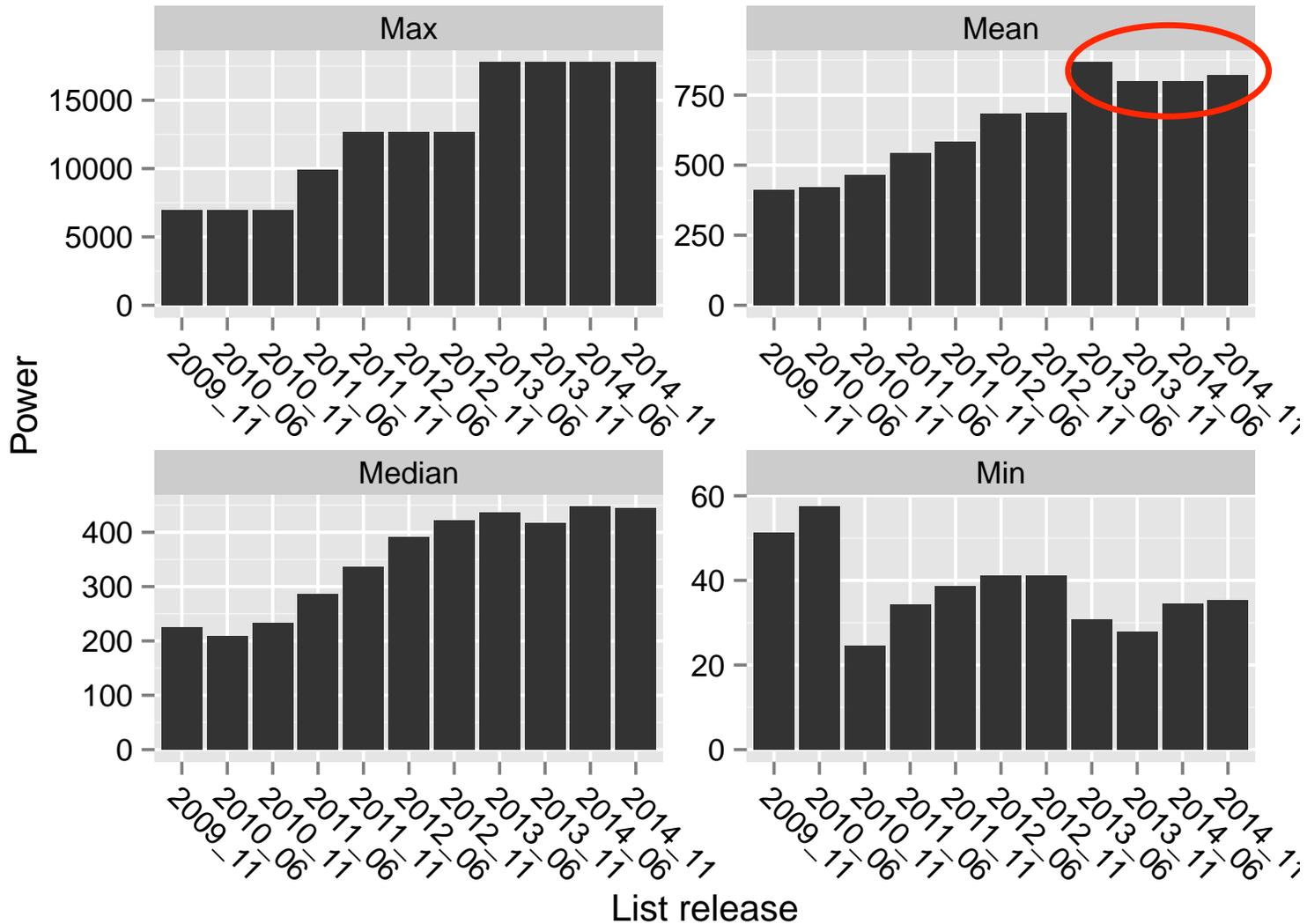
Trends: How Energy Efficient Are We?



Green500 Rank 1 10 100 500

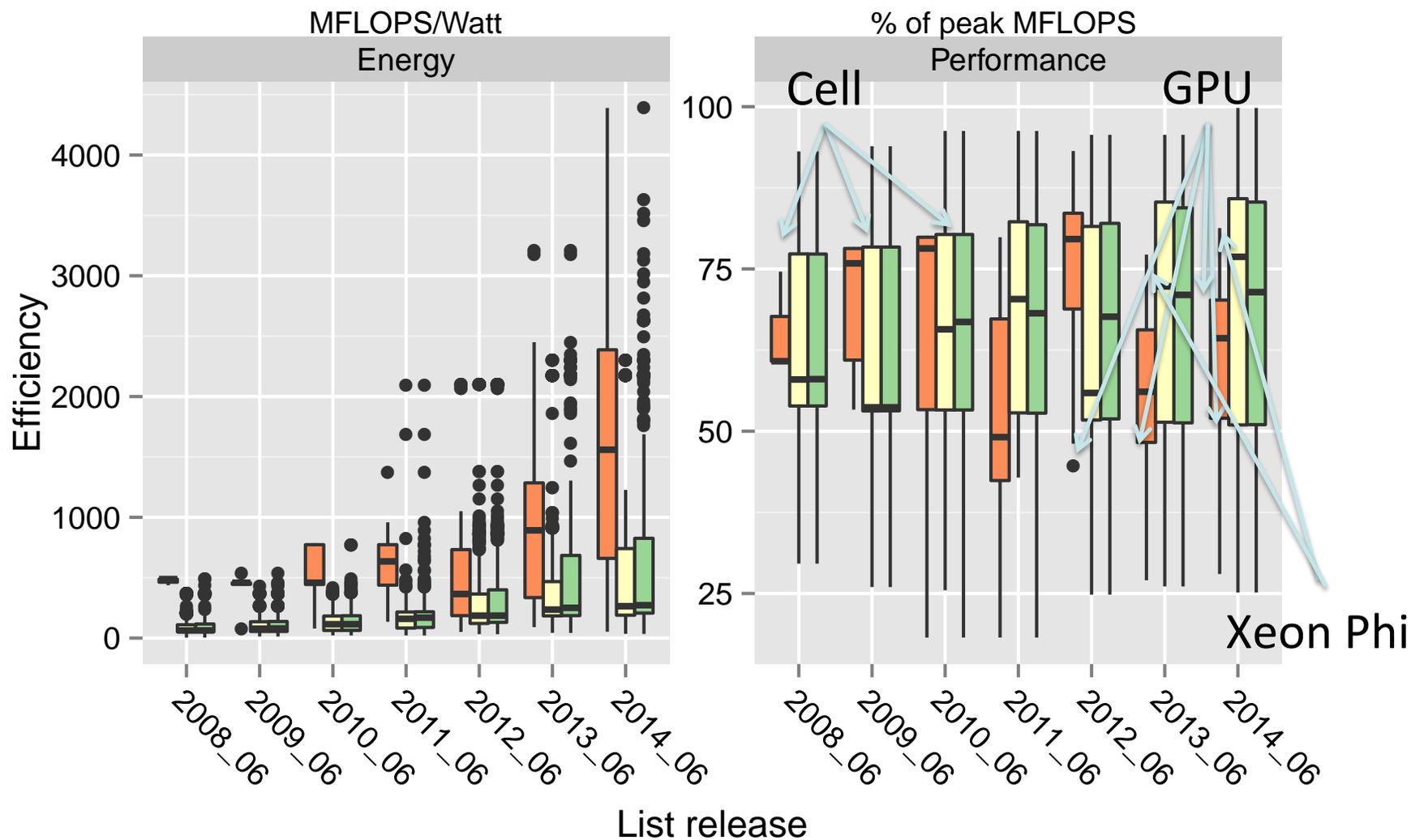


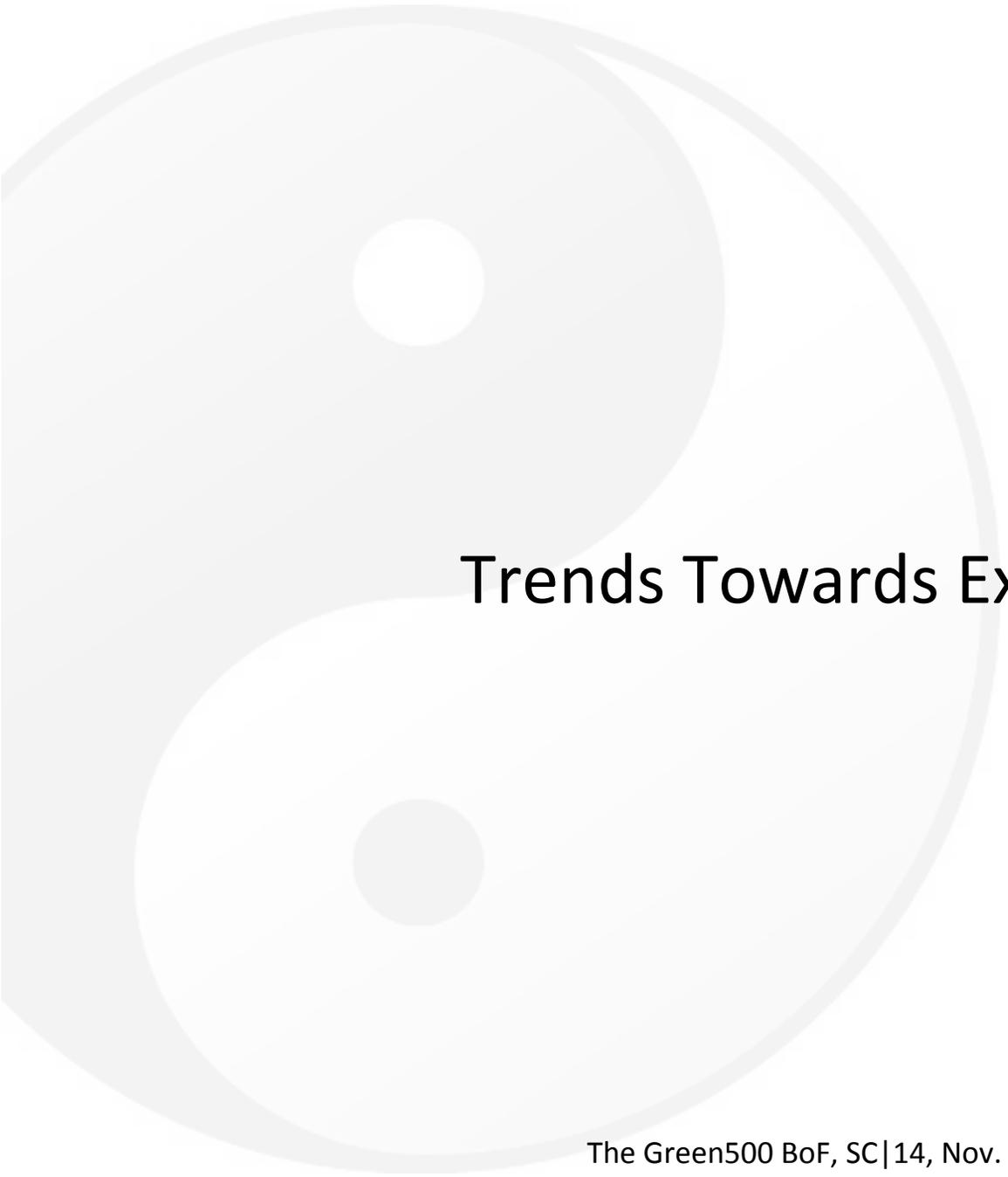
Trends in Power: Max, Mean, Median, Min



Trends: Energy vs Performance Efficiency

Machine type ■ Heterogeneous ■ Homogeneous ■ All





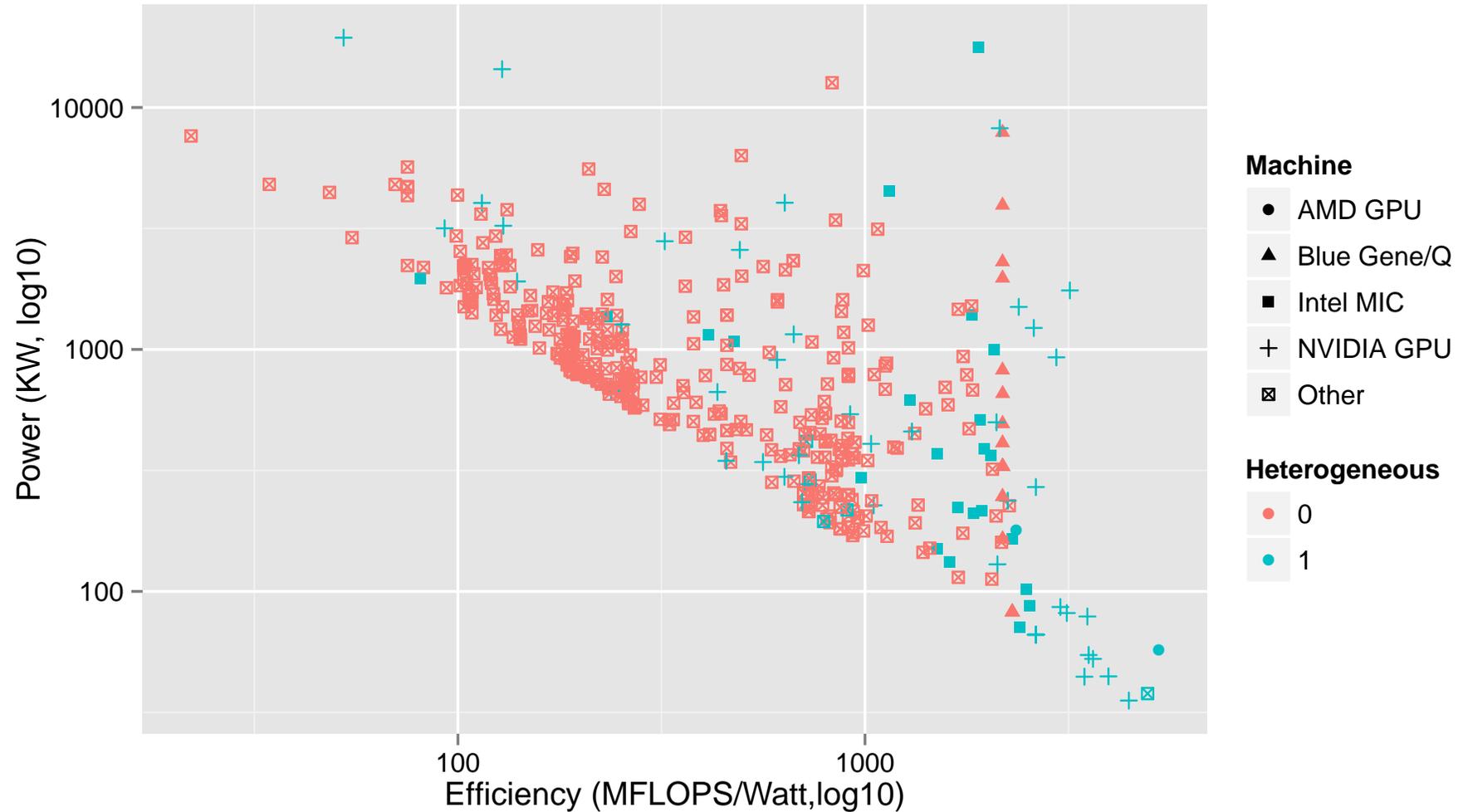
Trends Towards Exascale

The Green500 BoF, SC|14, Nov. 2014
POC: info@green500.org

Exascale Computing Study: Technology Challenges in Achieving Exascale Systems

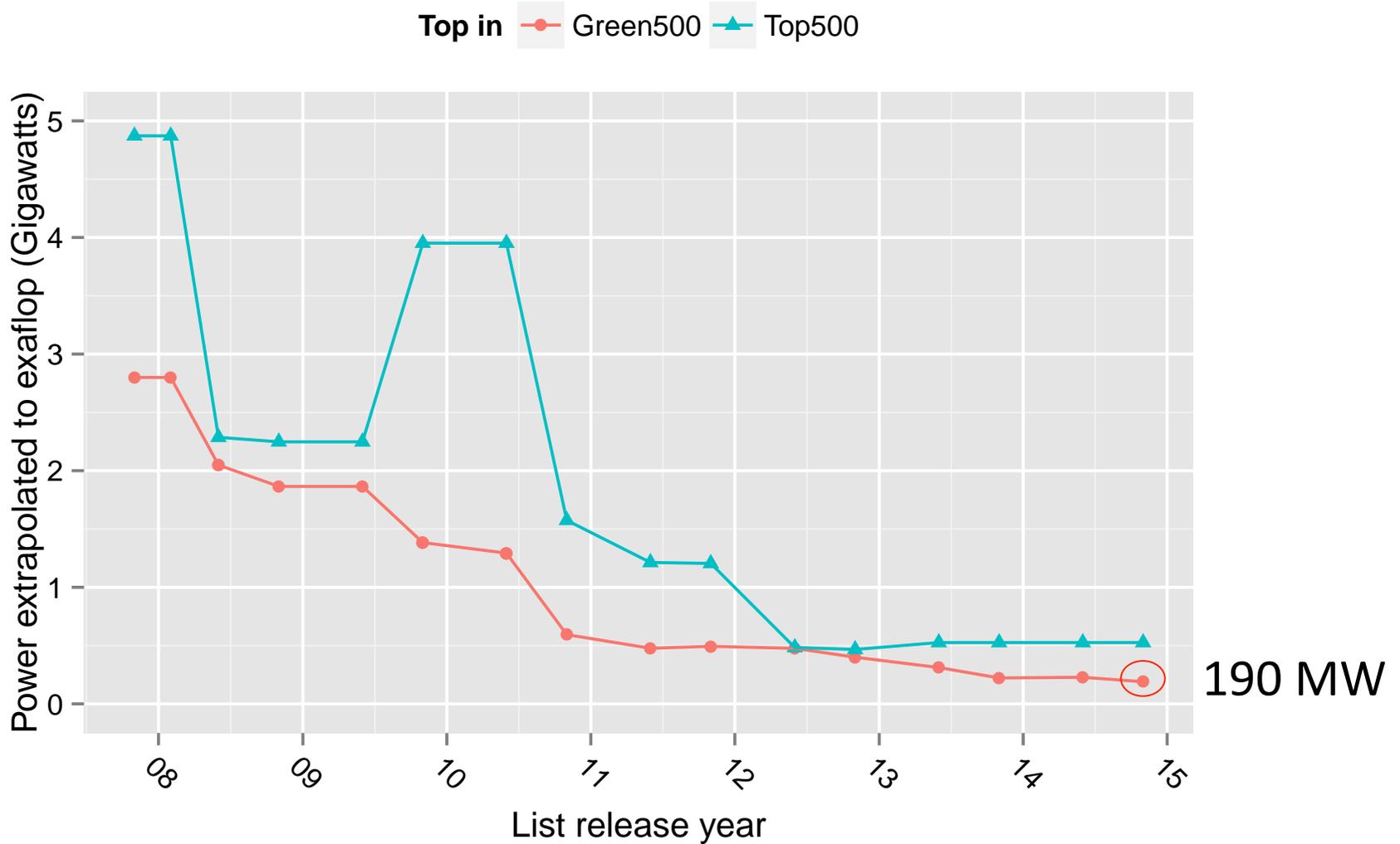
- Goal
 - “Because of the difficulty of achieving such physical constraints, the study was permitted to assume some growth, perhaps a factor of 2X, to something with a maximum limit of 500 racks and **20 MW** for the computational part of the 2015 system.”
- Realistic Projection?
 - “Assuming that Linpack performance will continue to be of at least passing significance to real Exascale applications, and that technology advances in fact proceed as they did in the last decade (both of which have been shown here to be of dubious validity), then [...] an Exaflop per second system is possible at around **67 MW**.”

Trends: Zoomed View of Nov. 2014 (By Machine Type)



The Green500 BoF, SC|14, Nov. 2014
POC: info@green500.org

Trends: Extrapolating to Exaflop



Evolution of the Green500

- *Methodologies for Measuring Power*
 - Collaboration between EE HPC WG, Green Grid, TOP500, and Green500, started in June 2011
- Research, Evaluation, Improvement, and Convergence on
 - Metrics
 - Methodologies
 - Workload

Evolution: Green500 Methodology

	Aspect 1: Time Fraction & Granularity	Aspect 2: Machine Fraction	Aspect 3: Subsystems Measured
Level 0	Derived numbers		
Level 1	20% of run: 1 average power measurement	(larger of) 1/64 of machine or 1kW	<p style="color: red;">Level 1+? 1/16 of machine + network</p> <ul style="list-style-type: none"> [Y] Compute nodes [] Interconnect net [] Storage [] Storage Network [] Login/Head nodes
Level 2	100% of run: at least 100 average power measurements	(Larger of) 1/8 of machine or 10kW	
Level 3	100% of run: at least 100 running total energy measurements	whole machine	

Where Are We Now?

- Feedback from the Community
 - Include cooling and associated infrastructure to the power
 - Focus on the energy efficiency of the machine itself
 - PUE: A measure of how efficiently a datacenter uses its power
 - Total Facility Power / IT Equipment Power
 - Software-based tuning for energy efficiency
 - Is it rewarding software innovation or gaming the system?
 - Phase-out Level 0 (via more reporting)
 - Why does it exist?
 - Incentivize non-reporting institutions to report.
 - Phasing out of Level 1 → Adoption of a Level 1+ or Level 2

Where Do We Want To Go?

- Continue building upon the momentum of the green HPC movement ...
 - Current Buy-In: ~ 60% of Green500 are submitted rather than derived (level 0) numbers
- Three aspects
 - Methodologies: Different levels of measurement methodologies
 - Level 0 and Level 1 → Level 2 and 3
 - Metrics: FLOPS/W → ???
 - Workloads: LINPACK → ???

Top 10 of the Green500

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	5,271.81	GSI Helmholtz Center	L-CSC - ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150 Level 1 measurement data available	57.15
2	4,945.63	High Energy Accelerator Research Organization /KEK	Suiren - ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC	37.83
3	4,447.58	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	35.39
4	3,962.73	Cray Inc.	Storm1 - Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m Level 3 measurement data available	44.54
5	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
6	3,543.32	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60
7	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray CS300 Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
8	3,459.46	SURFsara	Cartesius Accelerator Island - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4x FDR, Nvidia K40m	44.40
9	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
10	3,131.06	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41

TSUBAME-KFC (Kepler Fluid Cooling): An Ultra-Green Supercomputer Testbed in Tokyo Tech

**High Density
Compute Nodes with
Latest Accelerators**



**40 NEC/SMC 1U Servers
2 IvyBridge CPUs + 4 K20X GPUs
per node**

**Peak performance
217TFlops (DP)**

**GRC Oil-Submersion Rack
Processors 40~70°C
⇒ Oil <40°C**



Heat Exchanger

**Oil <40°C
⇒ Water <35°C**



**Heat Dissipation
to Outside Air**



Container

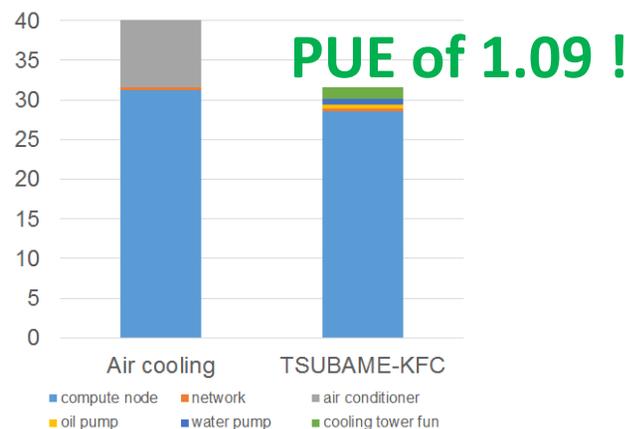
20 Feet Container (16m²)

Cooling Tower:

**Water <35°C
⇒ Outside**

Worlds' top efficiency, >4GFlops/W

- **Green #1 in Nov 13&Jun 14!**





This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.

GSIC Center, Tokyo Institute of Technology

is ranked

3rd

on the world's Green500 List of computer systems as of

November 2014



Wu-chun Feng, Co-Chair



Kirk Cameron, Co-Chair

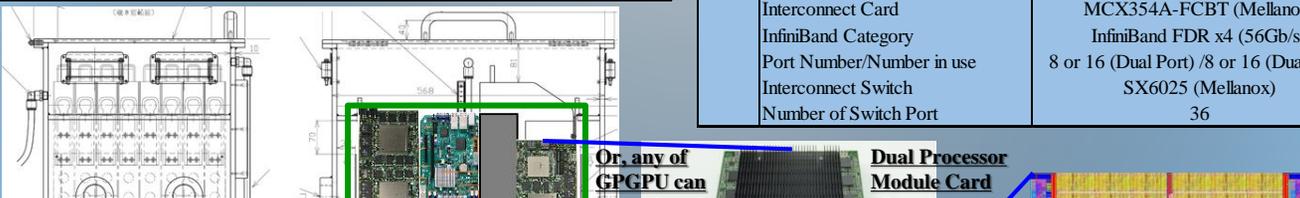
Submersion Liquid Cooling Supercomputer 「ExaScaler-1」

Ranked at # [] of Top500 with [] TFLOPS and Ranked at # [] of Green500 with [] GFLOPS/W
(Super 4 Tank Rack System at KEK Computing Research Center)

Submersion Liquid Cooling Tank Rack System 「ESLC-8」

ESLC-8 Submersion Cooling System (ExaScaler, Inc.)		
System Configuration		
Number of Unit	8U / Tank Rack	
Unit Motherboard	X9DRG-HTF (Supermicro)	
Cooling Method	Single Phase Direct Cooling	
Coolant / Coolant Volume	Fluorinert (3M) / 200L (approx.)	
Cooling Capacity	13,800kcal/h (at 25 Celcius Water)	
Pump Capacity	100L/min.	
GPU Board <i>ESLC-8 Tank Rack System is ready for NVIDIA/AMD/Intel GPGPUs.</i>		
Maximum GPU Board Length	Up to 280mm	
Number of GPU Boards per Unit	4	
Total Number of GPU Board	32	
Maximum Power Consumption	350 Watt per GPU Board	
Total Unit Power Capacity	1,800 Watt per Unit	
PCIe Interface	PCIe 3.0 x16 (32 Buses)	
Host System (Xeon E5-2xxx v3 with DDR4 configuration will be offered in Q2/15)		
Host System Processor	Xeon E5-2xxx v2 Family (Intel)	
Number of Processor Core	4 to 10	
Processor Speed	1.7 to 3.5GHz	
Number of Processor	16	
System Memory/Speed	DDR3/1,866MHz	
Total System Memory	up to 2TB (128GB per Xeon)	
Interconnect (Optional)		
Interconnect Card	MCX354A-FCBT (Mellanox)	
InfiniBand Category	InfiniBand FDR x4 (56Gb/s)	
Port Number/Number in use	8 or 16 (Dual Port) /8 or 16 (Dual Port)	

ExaScaler-1 Complete System (ExaScaler, Inc.)		
System Configuration		
Submersion Tank Rack System	ESLC-8 (ExaScaler, Inc.)	
Number of Unit	8U / Tank Rack	
Unit Motherboard	X9DRG-HTF (Supermicro)	
Cooling Method	Single Phase Direct Cooling	
Coolant	Fluorinert (3M)	
Cooling Capacity	13,800kcal/h (at 25 Celcius Water)	
Pump Capacity	100L/min.	
Main Manycore Processing System		
Super Manycore Processor	PEZY-SC (PEZY Computing, K.K.)	
Number of Processor Core	1,024	
Total Number of System Core	65,536	
Peak Procoessor Performance	1.5TFLOPS (Double) @733MHz	
Total System Peak Performance	96TFLOPS with 64 of PEZY-SC	
System Memory/Speed	DDR3/1,333MHz	
Total System Memory	2TB (32GB per PEZY-SC)	
PCIe Interface	PCIe 3.0 x16 (32 Buses)	
Host System		
Host System Processor	Xeon E5-2660v2 (Intel)	
Number of Processor Core	10	
Processor Speed	2.2GHz	
Number of Processor	16	
System Memory/Speed	DDR3/1,866MHz	
Total System Memory	2TB (128GB per Xeon)	
Interconnect		
Interconnect Card	MCX354A-FCBT (Mellanox)	
InfiniBand Category	InfiniBand FDR x4 (56Gb/s)	
Port Number/Number in use	8 or 16 (Dual Port) /8 or 16 (Dual Port)	
Interconnect Switch	SX6025 (Mellanox)	
Number of Switch Port	36	





This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.

High Energy Accelerator Research Organization KEK

is ranked

2nd

on the world's Green500 List of computer systems as of

November 2014



Wu-chun Feng, Co-Chair



Kirk Cameron, Co-Chair

The Lattice-CSC Cluster at GSI

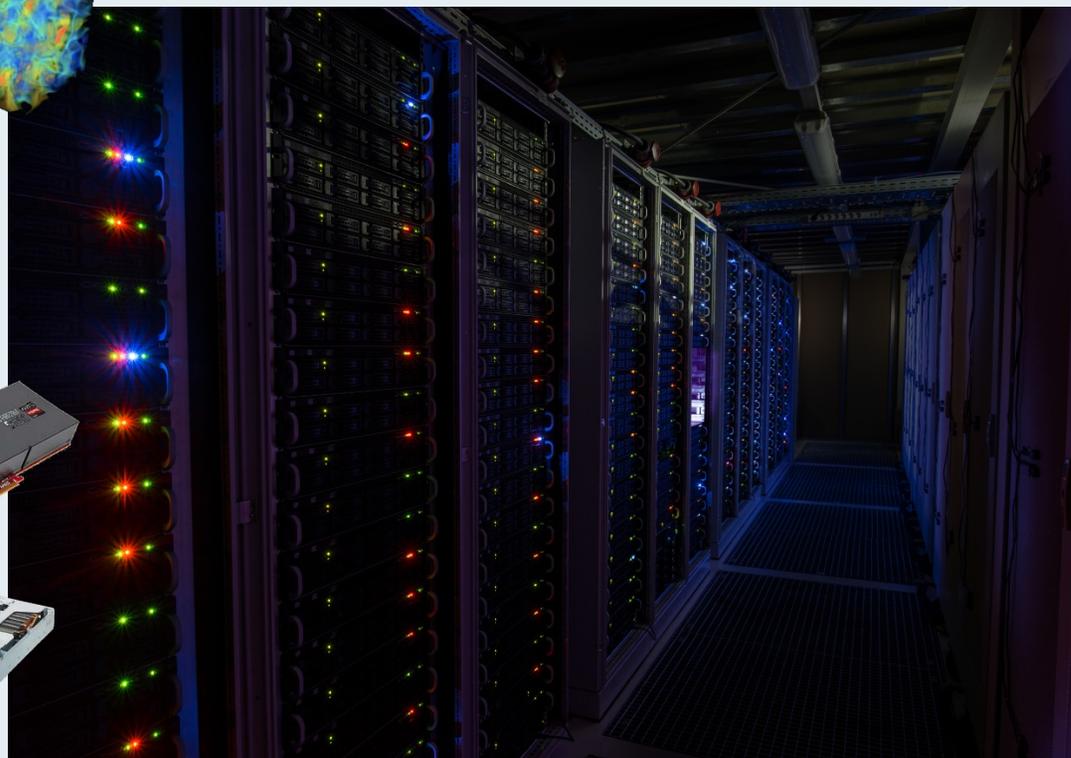
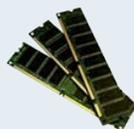
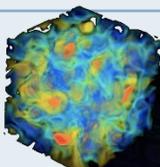
Lattice-CSC (at GSI):

- Built for Lattice-QCD simulations.
- Quantum Chromo Dynamics (QCD) is the physical theory describing the strong force.
- Very memory intensive.

160 Compute nodes:

- 4 * AMD FirePro S9150 GPU
- ASUS ESC4000 G2S Server
- 2 * Intel 10-core Ivy-Bridge CPU
- 256 GB DDR3-1600 1.35V
- FDR Infiniband

Installation ongoing, 56 nodes ready



Green DateCenter at GSI, Darmstat, Germany



This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.

GSI Helmholtz Center

is ranked

1st

on the world's Green500 List of computer systems as of

November 2014



Wu-chun Feng, Co-Chair



Kirk Cameron, Co-Chair

Acknowledgements

- Key Contributors
 - Balaji Subramaniam
 - Thomas Scogland
 - Vignesh Adhinarayanan
- **YOU!**
 - For your contributions in raising awareness in the energy efficiency of supercomputing systems